LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-235089

The Earth System Grid Center for Enabling Technologies: Focusing Technologies on Climate Datasets and Resource Needs

D. Williams

September 28, 2007

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The Earth System Grid Center for Enabling Technologies: Focusing Technologies on Climate Datasets and Resource Needs

1. Introduction (1 page)

Climate research is inherently a fundamentally multidisciplinary endeavour. As researchers strive to understand the complexity of our climate system they form multinational teams to tackle "Grand Challenge" problems. These multidisciplinary virtual organizations need a common software infrastructure to access the many large global climate model datasets and tools. It is critical that this infrastructure provide equal access to climate data, supercomputers, simulations, visualization software, whiteboard, and other resources. To this end, the Earth System Grid (ESG) Center for Enabling Technologies (ESG-CET) is a collaboration of U.S. research laboratories (ANL, LANL, LBNL, LLNL, NCAR, NOAA/PMEL, and ORNL) and a university (USC/ISI) working together to identify and implement key computational and informational technologies for advancing climate change science. Sponsored by the Department of Energy (DOE) Scientific Discovery through Advanced Computing (SciDAC)-2 program, through the Offices of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER), ESG-CET utilizes and develops computational resources, software, data management, and collaboration technologies to support observational and climate data archives.

In realizing our vision, our first steps began with "Prototyping an Earth System Grid" (ESG I). This project was initially funded under the DOE's Next Generation Internet (NGI) program, with follow-on support from BER and the Mathematical, Information, and Computational Sciences (MICS) office. In this prototype, we developed Data Grid technologies for managing the movement and replication of large datasets, and applied these technologies in a practical setting (i.e., an ESG-enabled data browser based on current climate data analysis tools), achieving cross-country transfer rates of more than 500 Mb/s. Having demonstrated the potential for remotely accessing and analyzing climate data located at sites across the U.S., we won the "Hottest Infrastructure" award in the Network Challenge event.

While the ESG I prototype project substantiated a proof of concept ("Turning Climate Datasets into Community Resources"), the SciDAC Earth System Grid (ESG) II project made this a reality. Our efforts targeted the development of metadata technologies (standard schema, XML metadata extraction based on netCDF, and a Metadata Catalog Service), security technologies (Web-based user registration and authentication, and community authorization), data transport technologies (GridFTP-enabled OPeNDAP-G for high-performance access, robust multiple file transport and integration with mass storage systems, and support for dataset aggregation and subsetting), as well as web portal technologies to provide interactive access to climate data holdings. At this point, the technology was in place and assembled, and ESG II was poised to make a substantial impact on the climate modelling community.

In 2004, the National Center for Atmospheric Research (NCAR), a premier climate science laboratory, began its first publication of climate model data into the ESG archives for restrictive use. Late that same year, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), an internationally recognized climate data center at Lawrence Livermore National Laboratory (LLNL), began its production service for climate model data germane to the Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report (AR4). (Because of international data requirements, restrictions, and timelines, the NCAR and PCMDI ESG data holdings were separated.) Since the release of these two institutional systems, ESG has become a world-renowned leader in developing technologies that provide scientists with virtual access to distributed data and resources.

As of late, over 7,000 registered users around the globe have downloaded more than 300 TB of data from aggregate ESG sites and over 300 scientific papers have been published, based on the analysis of the CMIP3 (IPCC AR4) data holdings. (See URL: http://www-pcmdi.llnl.gov/ipcc/subproject_publications.php.) In addition, ESG provides seamless access to over 190 TB of distributed climate simulation data.

In realizing our goal, not only is ESG leading the effort in the climate community towards standardization of material for the global federation of metadata, security, and data services required to analyze and access data worldwide, but helped to make major steps forward in the advancement of climate change research.

2. Overview of ESG (1 Page)

ESG is a large, production, distributed system – a DataGrid - with primary access points via three web portals: one for general climate research data, another dedicated to the IPCC activity, and a third for the Community Climate System Model (CCSM) Biogeochemistry (BGC) Working Group. Because of international data requirements, restrictions, and timeline, three separate web portals based on the same underlying integration of technologies were constructed to provide access to the vast data holdings made available through the aggregate ESG system. The next-generation ESG system will supersede these aggregate systems to provide ESG users with coherent access to ever-increasing rich and diverse collections of global community climate data.

Users of the ESG portal first undergo a registration process, where they are made known to the system and granted various privileges and access to data collections. The main portal page, shown in Figure 1, provides news, status, and live monitoring of the ESG. Once logged in, users may either search or browse ESG catalogs to locate desired datasets, with the option of browsing both collection-level and file/usage-level metadata. Based on this perusal of the catalogs, users may gather a collection of files into a "DataCart" or request an "aggregation," which allows them to request a specific set of variables subject to a spatiotemporal constraint. Selected data may then be downloaded to the user's system, including datasets that are on deep storage at multiple sites behind security firewalls. Group-based authorization mechanisms allow the ESG administrators to control which users can access which data. Behind all of this exists a collection of ESG management and data publishing tools, along with large-scale data transport tools.

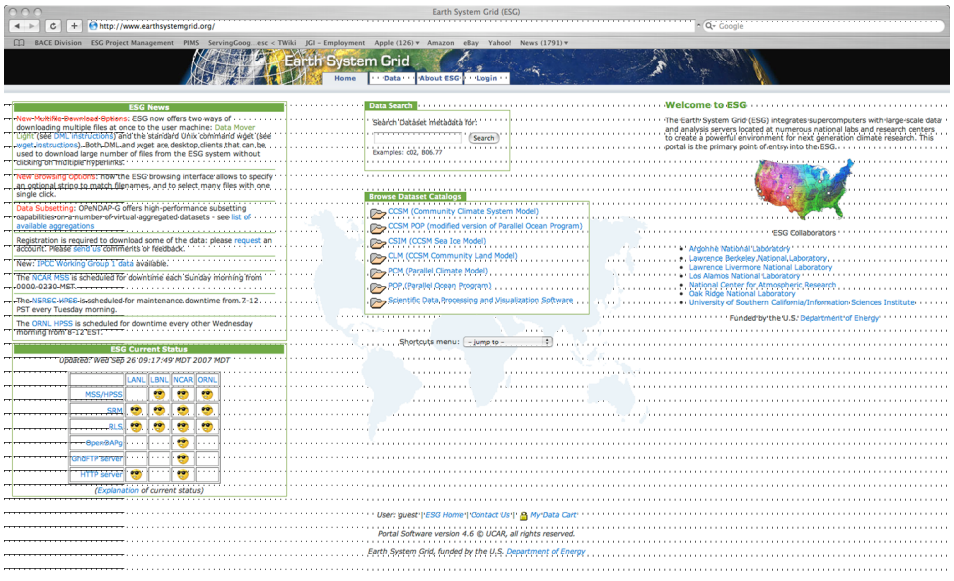

Figure 1: ESG Portal

The ESG system includes a metrics-gathering capability that keeps track of user activity. Interactive displays as well as reports allow us to track what data is downloaded, how often, and by whom. The resulting data has proved invaluable as a means of not only providing reporting to DOE and other groups on degree of use (its initial intent) but also as a means of guiding system development and optimization.

3. Overall Impact (1 Page)

ESG has had an influential impact upon the national and international climate community by enabling broad dissemination of important data holdings, including the Community Climate System Model (CCSM) data archive, the Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report (AR4) data archive, and now the BGC CCSM Carbon-Land Model Intercomparison Project (C-LAMP) data archive. All three archives are well known to the user community and since ESG's official release, the community has downloaded well over 300 TB of data (corresponding to over 1 million files) and has recorded a remarkable scientific publications list reaching well over 300 journal articles, all in a short time span.

As originally intended, the ESG team works closely with the CCSM community to publish CCSM model data into the ESG archives. Collaborating with CCSM scientists and data providers, the ESG team developed and utilized Grid technology that interfaces into the ESG metadata database allowing the CCSM community to completely view and manage all information related to generating, defining, and archiving CCSM model simulation runs. This interface, allows scientists to impose selective access control on the project runs, to sort information by any type, and to collaboratively enter data. The long-term goal is to tie the metadata ingestion process to the actual CCSM run workflow, so that model simulation metadata can be added automatically into the ESG data holdings. This user base represents climate scientists, analysts, educators, governments (both domestic and abroad), private industry, and many others. CCSM data (along with many other important datasets (e.g., the Parallel Climate Model (PCM) and the Parallel Ocean Program (POP)) accessed via ESG has been used in numerous scientific papers, impact analysis, urban planning, ecosystem monitoring, education, and other activities. By allowing access, ESG is enabling scientists, hardware and software engineers, universities and others to examine and learn how a state-of-the-art climate model works, and to provide suggestions and enhancements for its scientific accuracy, portability, and performance.

As an international challenge to the ESG system, ESG began its production service to distribute the IPCC / Working Group on Coupled Models (WGCM) data to the international climate community. The IPCC, which was jointly established by the World Meteorological Organisation (WMO) and the United Nations Environment Programme, carries out periodic assessments of the science of climate change. Fundamental to this effort is the production, collection and analysis of data from climate model simulations carried out by major international research centers. Analysis of a set of standard climate-change simulations from many modeling centers provides comprehensive understanding of the strengths and weaknesses of climate models. The IPCC and WGCM requested that PCMDI at LLNL collect model output data from these IPCC simulations, and distribute these to the community via ESG. Since this effort began, IPCC model runs published to the climate community via the CMIP3 (AR4 IPCC) ESG portal total to just over 35 TB (77,400 files), and some 1,400 users have registered to receive IPCC data for analysis. Overall, the number of files downloaded by the climate community totals over a million files, which is equivalent to over 300 TB of data. Figure 2 shows the daily download rate average at 500 GB. To date, the climate community has published some 320-research papers pertaining to analysis of the IPCC ESG data archive (see URL http://www-pcmdi.llnl.gov/ipcc/subproject_publications.php for citations).

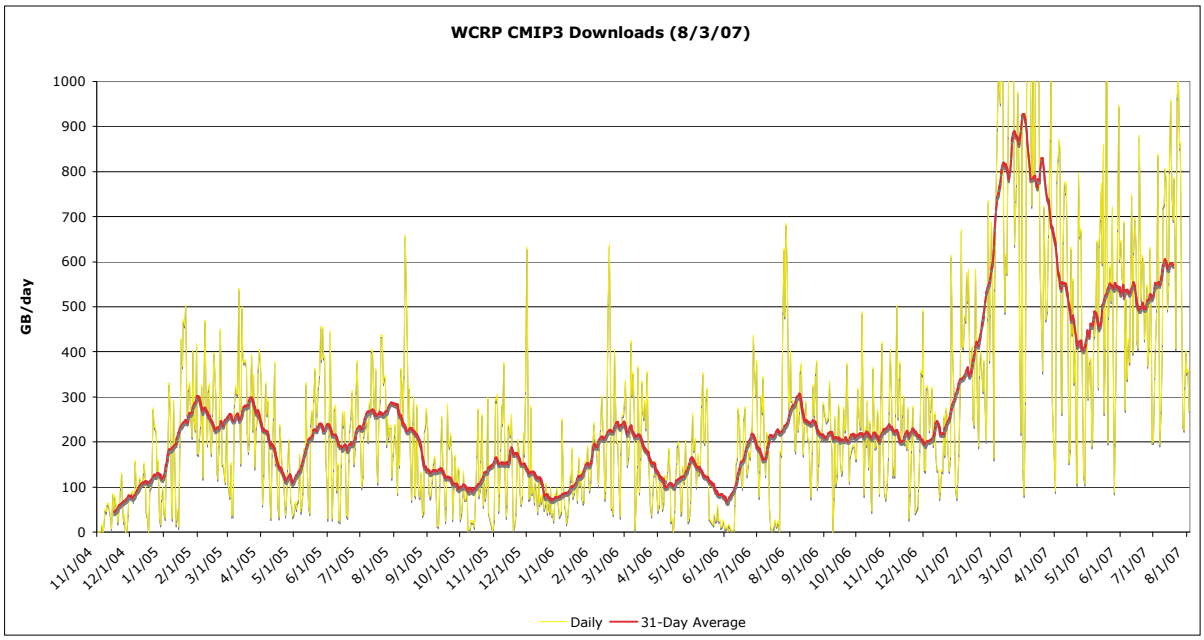

Figure 2: CMIP3 (IPCC AR4) Download Rates

New to ESG is the dissemination of BGC C-LAMP data. This model inter-comparison project has two terrestrial BGC modules linked to the same set of prescribed ocean BGC fluxes together with the CCSM's interactive atmosphere and interactive land surface modules. Two separate experiments envisioned for each terrestrial BGC module were decided for C-LAMP: one in which atmospheric data came from observations, the other in which it was calculated by CAM3, the current atmospheric component of the CCSM. The first experiment would determine how well land-air fluxes of CO2 are simulated given the observed climate; the second would determine the effect of the atmosphere model's climate bias (notably in precipitation) on the simulated CO2 fluxes. Visit the URL for more details on the inter-comparison experiments: http://climatemodeling.org/bgcmip. The C-LAMP experimental output is now being archived and disseminated on an ESG C-LAMP site modelled after the ESG CMIP3 (IPCC AR4) housed at LLNL. Noting the large data volume of C-LAMP model output, it is recommended that the BGC Working Group access and test the data only. Ultimately the working group will open up the data to anyone interested doing scientific research.

Knowledge and expertise gained from ESG have helped the climate community plan strategies to manage a rapidly growing data environment more effectively. Moreover, approaches and technologies developed under the ESG project have impacted data-simulation integration in other disciplines, such as astrophysics, molecular biology and materials science.

4. The Next-Generation ESG (1 Page)

Building upon the years of ESG success, the ESG-CET is developing a next-generation environment targeted at enabling flexible, efficient, universal access to even larger datasets and distributed worldwide resources for the purpose of advancing climate and related impacts research and assessment. In creating this new community infrastructure, ESG-CET will turn even more climate model data into true community resources and place-advanced capabilities into the hands of a substantial user community base. Our high-level goals are driven by scientific objectives relevant to DOE scientific priorities over the next several years. In brief, they are:

- Sustain successful existing ESG services.

- Address scientific needs related to projections of data management and analysis requirements, with a particular focus on:
    - Preparing for the CMIP4 IPCC 5th Assessment Report (AR5) in 2010.
    - Publishing and enabling processing of the massive data produced by the *Climate Science Computational End Station (CCES)* at ORNL's NCCS/LCF.
    - Support a wide-range of climate model evaluation activities aimed at improving climate change research.

To support this effort, we will broaden ESG to support multiple types of model and observational data, provide more powerful (client-side) ESG access and analysis services, enhance interoperability between common climate analysis tools and ESG, and enable end-to-end simulation and analysis workflow. Figure 3 depicts the scientific data management and analysis requirements in relationship to the ESG development timeframe. We specifically note that *a distributed testbed for IPCC AR5 must be in place by early 2009.*

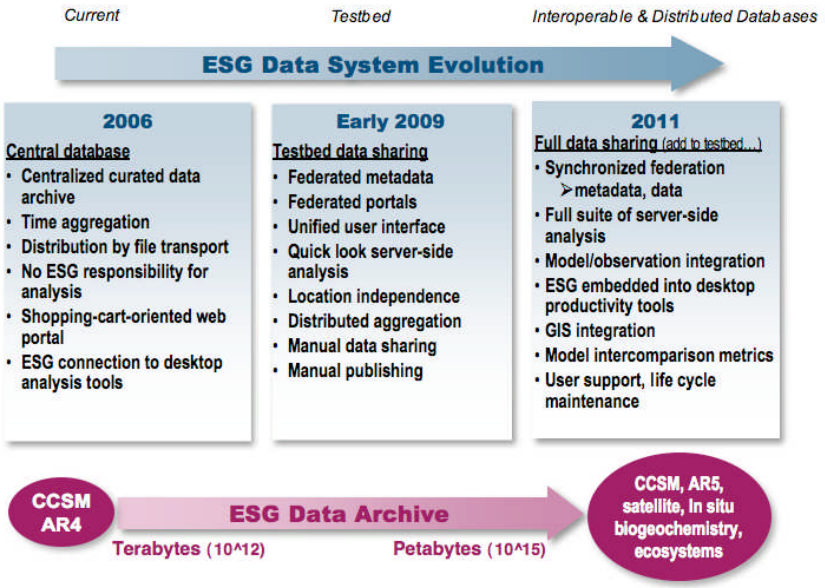

Figure 3: Evolving ESG to the Petascale: High-level ESG-CET Roadmap

5. Conclusion (1/2 Page)

The Earth System Grid was initially developed and deployed under the SciDAC-1 program to meet the needs of the DOE's Climate Change Prediction Program to disseminate data to its research community. It also supported the international IPCC AR4 activity. Under SciDAC-2, the Earth System Grid Center for Enabling Technologies (ESG-CET) is carrying this work forward to meet the formidable challenges of the next phase of climate change research. Development efforts are underway to enable the deployment of a distributed infrastructure that is global in scope and that supports petascale data volumes, more complex models and data, semantically based user interfaces, and a broader range of services, along with a variety of analysis and visualization applications.